Experimental Study of Aggressive Undervolting in FPGAs

Behzad Salami, Osman S. Unsal, Adrian Cristal Kestelman

Barcelona Supercomputing Center (BSC), Barcelona, Spain.

{behzad.salami, Osman.unsal, adrian.cristal}@bsc.es

***Keywords*— FPGAs, Voltage Underscaling, Energy, Reliability**

EXTENDED ABSTRACT

In this work, we evaluate aggressive undervolting, i.e., voltage scaling below the nominal level to reduce the energy consumption of Field Programmable Gate Arrays (FPGAs). Usually, voltage guardbands are added by chip vendors to ensure the worst-case process and environmental scenarios. Through experimenting on several FPGA architectures, we measure this voltage guardband to be on average 39% of the nominal level, which in turn, delivers more than an order of magnitude power savings. However, further undervolting below the voltage guardband may cause reliability issues as the result of the circuit delay increase, i.e., start to appear faults. We extensively characterize the behavior of these faults in terms of the rate, location, type, as well as sensitivity to environmental temperature, with a concentration of on-chip memories, or Block RAMs (BRAMs). Finally, we evaluate a typical FPGA-based Neural Network (NN) accelerator under low-voltage BRAM operations. In consequence, the substantial NN energy savings come with the cost of NN accuracy loss. To attain power savings without NN accuracy loss, we propose a novel technique that relies on the deterministic behavior of undervolting faults and can limit the accuracy loss to 0.1% without any timing-slack overhead.

A. Introduction

The power consumption of digital circuits, e.g., FPGAs, is directly related to their operating supply voltages. On the other hand, usually, chip vendors introduce a conservative voltage guardband below the standard nominal level to ensure the correct functionality of the design in the worst-case process and environmental scenarios. For instance, this voltage guardband is empirically measured to be 12%, 20%, and 16% of the nominal level in commercial CPUs [1], GPUs [2], and DRAMs [3], respectively. However, in many real-world applications, this guardband is extremely conservative and eliminating it can result in significant power savings without any overhead. Motivated by these studies, we extended the undevolting technique to commercial FPGAs, with a preliminary concentration on on-chip memories, or Block RAMs (BRAMs). Our experiments cover several representative FPGA platforms from Xilinx, a main vendor including a VC707, two identical samples of KC705, and a ZC705. The experimental results show the voltage guardband to be on average 39% of the nominal level (Vnom= 1V, Vmin = 0.61V), which in turn, directly delivers an order of magnitude BRAM power savings. Further undervolting below the minimum safe voltage, i.e., Vmin delivers more power savings up to 40%; however, causes faults occurrence in some locations of some of BRAMs. These faults are the consequence of timing violations since the circuit delay increases by further undervolting. Note that simultaneously downscaling the frequency is a promising approach to prevent the generation of these faults; however, it can limit the energy reduction achievement. Alternatively, our aim is to understand the behavior of these faults, through which customized and low-overhead fault mitigation techniques can be deployed to achieve power saving gains.

B. Experimental Methodology

The overall methodology is shown in Fig. 1. Our FPGA design includes raw Read/Write accesses to BRAMs, while their supply voltage, i.e., VCCBRAM is controlled in the host through the Power Management Bus (PMBus) interface. The onboard voltage regulator with the part number of UCD9248 has the responsibility to handle these PMBus commands and set the appropriate voltage to different components, e.g., BRAMs. Note that other FPGA components, e.g., LUTs, DSPs, operate at their default nominal voltage levels.

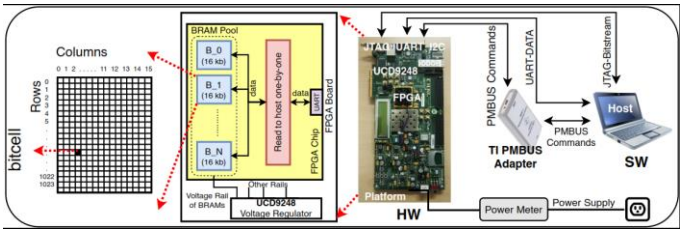


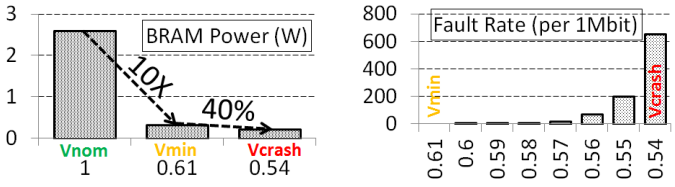
Fig. 1 The overall methodology for FPGA undervolting experimental study.

C. Selected Results

Through experiments on this setup, the overall power and reliability trade-off is summarized in Fig.2 for VC707, when VCCBRAM is downscaled from the nominal level Vnom = 1V to the minimum level that the FPGA practically operates, Vcrash = 0.54V. As can be seen, BRAMs start experiencing faults in regions below Vmin = 0.61V with an exponentially increasing behavior up to 653 faults per 1Mbit equals to 0.06%. Major observed properties of these faults are summarized as follows:

- There is significant variability of fault rate among different BRAMs, which is the consequence of the inherent process variation. Through our experiments, we observed that more than 38.9% of BRAMs never experience faults. Also, among BRAMs the maximum, minimum, and average fault rate are 2.84%, 0%, and 0.06%.
- The faults locations and rate do not change over time. Also, by experimentally evaluating different data patterns, we observed that the fault rate directly depends on the number of ”1” bits since a vast majority of generated faults are ’1’ to ’0’ bit-flips.
- More than 90% of undervolting faults are single-bit, and a further 7% are double-bit faults. Due to this observation on the behavior of faults, the built-in ECC of BRAMs can be effective to mitigate these faults. Note that the built-in ECC of BRAMs has the type of Single-Error Correction and Double-Error Detection (SECDED) capability, potentially with good efficiency to mitigate BRAMs undervolting faults.
- Faulty bitcells in a certain voltage stay faulty in lower voltages, as well, and potentially, expand to other bitcells. This property is called Fault Inclusion Property (FIP). Our work experimentally confirms that FIP exists in

FPGAs, under aggressive low-voltage operations. FIP can be potentially used to build efficient fault mitigation techniques.



(a) Power Saving Gain (b) Reliability Cost

Fig. 2 Power saving gain and reliability costs of FPGA BRAMs through undervolting.

D. CASE-STUDY: NEURAL NETWORK (NN)

In this section, we present and discuss results of our study on the impact of BRAMs undervolting in a typical FPGA-based NN accelerator. When VCCBRAM is underscaled in the critical region between Vmin= 0.61V until Vcrash= 0.54V, faults occurring in some of bitcells degrades the NN accuracy. Hence, the classification error is increased from 2.56% (inherent classification error without any fault) to 6.15% when VCCBRAM= Vcrash= 0.54V, see Fig. 3. The NN classification error (left y-axis) increases exponentially, correlated directly with the fault rate increase in BRAMs (right y-axis), as expected.

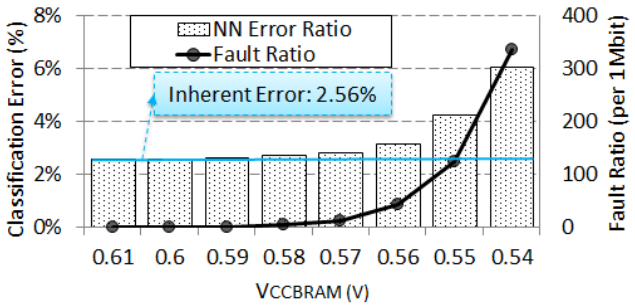


Fig. 3 Impact of BRAMs voltage scaling in the NN classification error, lowering VCCBRAM from Vmin = 0.61V to Vcrash = 0.54V (VC707).

E. Conclusion and Future Enhancement

This paper experimentally evaluated the supply voltage underscaling below the nominal level in commercial FPGAs. We discovered that there is a significant voltage guardband gap, where data can be safely retrieved from BRAMs. However, by further undervolting observable faults occur, as a result of the timing delay increase. We extensively characterized the behavior of these faults, more specifically for on-chip memories of FPGAs. Finally, we evaluated the impact of the undervolting in the accuracy and power of an FPGA-based NN accelerator in the inference phase. To attain the power efficiency without NN accuracy loss, we proposed an efficient application-aware BRAM placement algorithm that relies on the behavior of undervolting faults. As an ongoing work, we are working on a more comprehensive voltage scaling in other components of commercial FPGAs and on different FPGA technologies of vendors.

F. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union's Horizon 2020 Programme under the LEGaTO Project (www.legato-project.eu), grant agreement n◦ 780681.

References

[1] A. Bacha and R. Teodorescu. "Dynamic reduction of voltage margins by leveraging on-chip ecc in itanium ii processors," In ACM SIGARCH Computer Architecture News, volume 41, pages 297–307. ACM, 2013.

[2] J. Leng, A. Buyuktosunoglu, R. Bertran, P. Bose, and V. J. Reddi. "Safe limits on voltage reduction efficiency in gpus: a direct measurement approach", In 2015 48th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), pages 294–307. IEEE, 2015.

[3] K. K. Chang, G. Yaglıkc, S. Ghose, A. Agrawal, N. Chatterjee, A. Kashyap, D. Lee, M. O'Connor, H. Hassan, and O. Mutlu. "Understanding reduced-voltage operation in modern dram devices: Experimental characterization, analysis, and mechanisms", In Proceedings of the ACM on Measurement and Analysis of Computing Systems, 1(1):10, 2017.

Author biography

Behzad Salami is a post-doc researcher in the Computer Science (CS) department of Barcelona Supercomputing Center (BSC). He received his Ph.D. in Computer Architecture from Universitat Politècnica de Catalunya (UPC) in 2018. During his Ph.D. studies, he visited the University of Manchester (UNIMAN) as a research internship student using a collaboration grant awarded from HiPEAC. Also, he obtained MS and BS degrees in Computer Engineering from Amirkabir University of Technology (AUT) and Iran University of Science and Technology (IUST), respectively. He was/is involved in several H2020/FP7 EU-funded research projects including AXLE (Advanced Analytics for Extremely Large European Databases), LEGaTO (Low Energy Toolset for Heterogeneous Computing), and EuroEXA. His research interests are heterogeneous computing and low-power & fault-resilient hardware accelerators.